Int. J. on Recent Trends in Engineering and Technology, Vol. 8, No. 1, Jan 2013

"Performance Evaluation of Phish Mail Guard: Phishing Mail Detection Technique by using Textual and URL analysis"

Jayshree Hajgude[1] , Dr. Lata Ragha[2]
[1]V.E.S. Institute of Technology / Information Technology, Mumbai ,India.
Email:jayshreehajgude@gmail.com
[2]Terna Engineering College/Computer Engineering ,Mumbai, India
Email:Latha.ragha@gmail.com

***Abstract* — Phishing emails usually contain a message from a credible looking source requesting a user to click a link to a website where user is asked to enter a password or other confidential information. Most phishing emails aim at withdrawing money from financial institutions or getting access to private information. Phishing has increased enormously over the last years and is a serious threat to global security and economy. Phishing attacks are becoming more frequent and sophisticated. There are a number of possible countermeasures to phishing. A number of anti-phishing solutions have been proposed to date. Some approaches attempt to solve the phishing problem at the e-mail level. A technique must be capable of determining whether an email is legitimate or a phishing, given only the URL and the email content. URL and textual content analysis of email will results in a highly accurate anti phishing email classifier. We proposed a technique where we considered the advantages of blacklist, white list and heuristic technique for increasing accuracy and reducing false positive rate. In heuristic technique we are using textual analysis and URL analysis of e-mail. Since most of the phishing mails have similar contents, our proposed method increased the performance by analyzing textual contents of mail and lexical URL analysis. This technique detect phishing mail if DNS in actual link is present in blacklist.DNS is present in white list then it is considered as legitimate DNS. If it is not present in blacklist as well as white list then it is analyzed by comparing senders DNS and DNS exists in link. This method analyzes URL with the help of lexical features of URL. Contents of mails are also analyzed because most of the phishing mail has similar contents. With the help blacklist and white list we are avoiding detection time for phishing and legitimate email. At the same time we are decreasing false positive rate by combining features of DNS, textual content analysis of email and URL analysis.**

***Index Terms* - URL, DNS, Phishing, lexical.**

I. Introduction

Phishing is a type of attack where the attacker creates a replica of an existing Web page to fool users into submitting personal, ûnancial, or any other sensitive data like password data to what they think is their service provider's Website. Phishing is the criminally fraudulent process of attempting to acquire sensitive information such as usernames, passwords and credit card details by masquerading as a legitimate trusted by customers in an electronic communication. Communications purporting to be from banks, online organizations, internet services providers, online retailers, and insurance agencies and so on. Phishing is typically carried out by email, and it often directs users to enter details at a fake website which is almost identical to the legitimate one. Phishing is the process of fooling a consumer into divulging personal information, such as credit card numbers or passwords, usually by sending an email carefully constructed to appear as if it's from a bank or other trusted entity, such as PayPal. As people increasingly rely on the Internet for business, personal finance and investment Internet fraud takes many forms, from phony items offered for sale on eBay, to scurrilous rumors that manipulate stock prices, to scams that promise great riches if the victim will help a foreign financial transaction through his own bank account. One interesting species of Internet fraud is phishing. Phishing attacks use email messages and web sites designed to look as if they come from a known and legitimate organization, in order to deceive users into disclosing personal, financial, or computer account information. Phishing emails usually contain a message from a credible looking source requesting a user to click a link to a website.

However, phishing has become more and more complicated and sophisticated so that phishers can bypass the filter set by current anti-phishing techniques and cast their bait to customers and organizations. A possible solution is to create a robust classifier to enhance the phishing email detection and protect customers from getting such emails. By analyzing phishing emails, it is observed that phishing emails often include certain phrases, for example, "security", "verify your account", "if you don't update your details within 2 days, your account will be closed", "click here to access to your account" and so on. These phrases may appear in the "subject:" line in an email or email content. Therefore, most phishing emails are largely similar in wording, especially the most important terms, such as "security", "expire", "unauthorized", "account", "login", etc. Such terms are useful to classify if an email is a phishing email [1][2][3]. In addition, Phishing emails often alert customer to click links to other websites which the real link is not the same as it is shown in the pages. In the proposed method we are using hybrid method for phishing mail detection which is a combination of blacklist, white list and heuristic technique. In heuristic

DOI: 01.IJRTET.8.1.61

technique we are considering textual and URL analysis for further classification. Hybrid email classification is used to enhance the classification accuracy of email messages. A number of features are extracted from email messages like text content, DNS name from visible link, URL features. This result into representing each message as a set of values where each value shows existence of that feature in that e-mail.

In this paper section II describes background and related work on Phishing mail detection methods and their drawbacks. Section III describes proposed phishing mail detection technique. Section IV deals performance analysis and section V includes performance comparison. Section VI contains conclusion.

II. Background And Related Work

In this paper, we assume that phishers use e-mail as their major method to carry out phishing attacks.

A. Procedure of Phishing Attacks

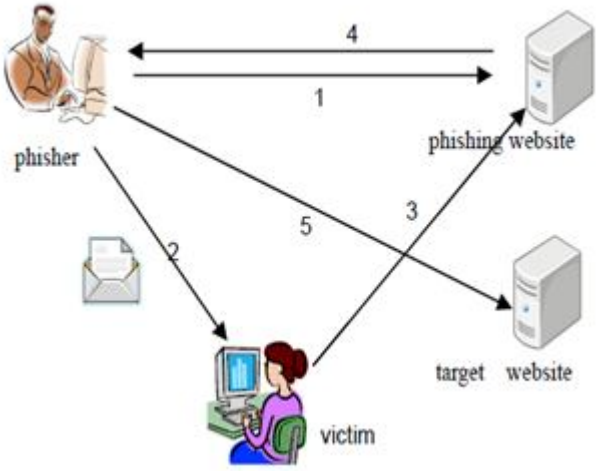

Figure 1. Procedure of Phishing Attacks

Phishing attack procedure is depicted in figure 1. Following steps are involved in phishing attack

1. Phishers set up a counterfeited Web site which looks exactly like the legitimate Web site, including setting up the web server, applying the DNS server name, and creating the web pages similar to the destination Website, etc.
2. Send large amount of spoofed e-mails to target users in the name of those legitimate companies and organizations, trying to convince the potential victims to visit their Web sites.
3. Receivers receive the e-mail, open it, click the spoofed hyperlink in the e-mail, and input the required information.
4. The confidential information is transmitted from a phishing server to the phisher.
5. Phishers steal the personal information and perform their fraud such as transferring money from the victims.

The phisher uses the identity information of the victim to the target website and impersonates the victim's identity to gain the illegal financial benefits. The e-mail directs the user to visit a Web site where they are asked to update personal information, such as passwords, credit card details and bank account numbers.

B. Types of phishing attack techniques

Figure 2 shows different phishing attack techniques

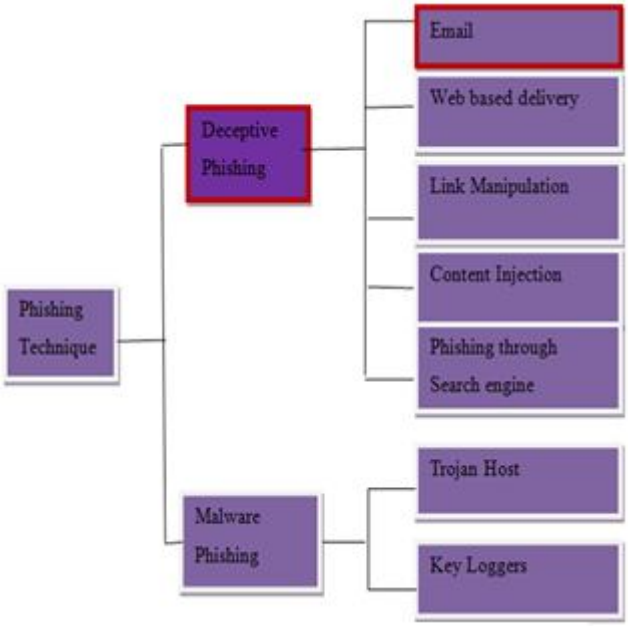

Figure 2. Different types of Phishing attack

Two different types of phishing attacks may be distinguished [4].

1. Malware-based phishing

For malware-based phishing is a malicious software is spread by deceptive emails or by exploiting security holes of the computer software and installed on the user's machine. Then the malware may capture user input, and confidential information may be sent to the phisher.

2. Deceptive phishing.

Deceptive phishing, in which a phisher sends out deceptive emails pretending to come from a reputable institution, e.g. a bank. In general, the phisher urges the user to click a link to a fraudulent site where the user is asked to reveal private information, e.g. passwords. This information is exploited by the phisher, e.g. by withdrawing money from the users bank account. Deceptive phishing Phishers may send the same email to millions of users, requesting them to fill in personal details. These details will be used by the phishers for their illegal activities.

C. Phishing mail detection Techniques

Figure 3 shows different types of phishing detection techniques. Blacklists and heuristic are arguably the most popular phishing detection techniques. As evaluated in, although blacklists achieve low false positives, their detection rates suffer at zero-hours and are evaluated to detect only 20% of zero-hour phishing attacks. On the other hand, heuristics are able to constantly detect phishing attacks at a similar rate. However, heuristics were evaluated to have high

DOI: 01.IJRTET.8.1.61

false positives.

Effectiveness of a blacklist-based solution depends on the time it takes until a phishing site is included. This is because many phishing pages are short-lived and most of the damage is done in the time span. The techniques are described in detail below .The suspicious URL is matched against a list of known Phishing sites. This method is susceptible to "zero day attacks". Also, techniques like URL obfuscation and routing through alternate domain name can hinder this method ineffective.

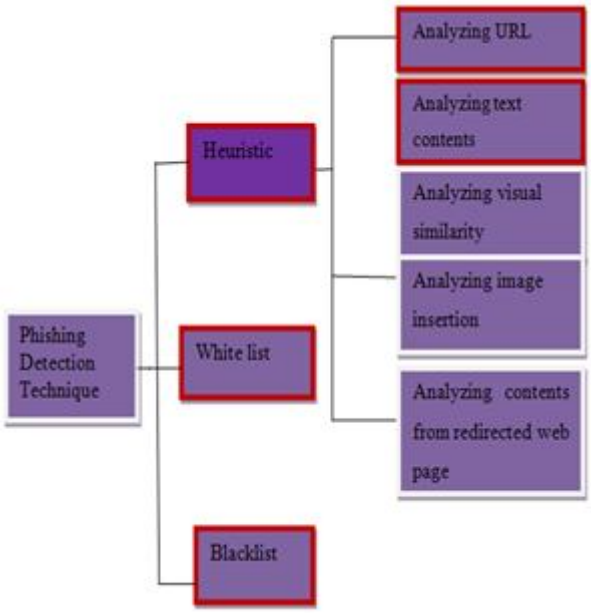

Figure 3. Phishing Mailing Detection Techniques

Most of the heuristics used are subjective and produce a large number of false positives. This solution is not limited to URL processing, but also analyzes the page layout. Although some heuristics are used in this solution, they are used only in the pre-processing stages, and the actual phish detection is completely independent of them. Drawbacks of this method are as mentioned below:

D. Literature Survey

There were lots of methods proposed for phishing mail detection .However, false positives have been observed in these methods. Also, a web site routed through content distribution network would create problems for domain based checks. As evaluated in [5][6], although blacklists achieve low false positives, their detection rates suffer at zero-hours and are evaluated to. detect only 20% of zero-hour phishing attacks. On the other hand, heuristics are able to constantly detect phishing attacks at a similar rate. However, heuristics were evaluated to have high false positives.

Phish Block is a hybrid phishing technique which is a combination of blacklist and heuristic approach and is explained in [7]. Lookup based systems suffer from high false negatives while classifier systems suffer from high false positives. To better detect fraudulent websites, we propose in this work an efficient hybrid system that is based on both lookup and a support vector machine classifier that checks features derived from websites URL, text and linkage this method is very complex and tested for small dataset.Next technique introduced was PhishCatch. PhishCatch is a heuristic based algorithm which will detect phishing emails and alert the users about the phishing emails [8]. The phishing filters and rules in the algorithm are formulated after extensive research of phishing methodologies and tactics. Phish catch rate of this technique is less.

John Yearwood, Musa Mammadov and Arunava Banerjee have proposed a heuristic method called profiling phishing email based on hyperlink information. This technique uses hyperlinks in the phishing emails as features and structural properties of emails along with whois information on hyperlinks as profile classes. But drawbacks of this method are for blacklisted url it is time consuming.No valid criterion for measuring the importance of the classes present in profiling [9].

PILLER is a machine learning based approach to e-mail classification [10]. The tool decides that whether some communication is deceptive, that is whether it is designed to trick the user into believing they are communicating with a trusted source, when in reality the communication is from an attacker. The decision is based on information from within the email or feature vector itself combined with information from external sources.

Bergholz, De Beer, Glahn, Moens, Gerhard and Strobel proposed a Machine Learning classifier with model-based features that is, features that themselves are classification models and require to be trained first prior to their use by a parent classifier [11]. The proposed classifier used a total of 27 features, two of which were model-based features.

Jeong-Ho Chang proposed a technique called Improved Phishing Detection using Model-Based Features [12]. This technique uses heuristic technique for phishing mail detection. But drawbacks of this method are low accuracy and blacklist not considered.

Chandrasekaran proposed a technique to classify phishing based on structural properties of phishing emails [13].They have used a total of 25 features mixed between style markers (e.g. the words suspended, account, and security) and structural attributes, such as the structure of the subject line of the email and the structure of the greeting in the body.

Lexical URL Analysis for Discriminating Phishing and Legitimate E-Mail Messages is proposed in [14]. In proposed method we are try to minimize false positive rate. Andrew Jones proposed Lexical URL Analysis for Discriminating Phishing and Legitimate E-Mail messages. The center claim of this paper is that lexical URL analysis technique can enhance the classification accuracy of email classifiers. We proposed and implemented a phishing mail detection technique phish mail guard. Phish mail guard is combination of blacklist, white list and heuristic approach.

DOI: 01.IJRTET.8.1.61

III. Phishing Mail Detection

A. Worlflow of Phishing Mail Detection

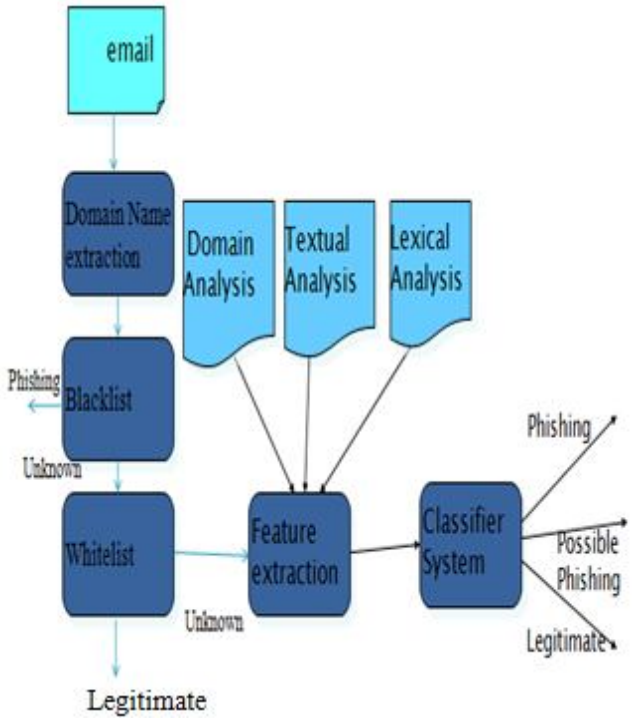

Figure 4. Work flow of Phishing Mail Detection

Work flow of proposed method is shown in figure 4.In the proposed method we are using hybrid method for phishing mail detection which is a combination of blacklist, white list and heuristic technique. In heuristic technique we are considering textual and lexical URL analysis for further classification. Hybrid email classification is used to enhance the classification accuracy of email messages. A number of features are extracted from email messages like text content, DNS name from visible link, URL features.

B. Modules used in Phishing Mail Detection

Proposed method mainly includes three Modules DNS analyzer, Classifier system, Lookup System. DNS analyzer component checks e-mail is phishing or not phishing by analyzing visual DNS and actual DNS in e-mail. This module checks the DNS of hyperlink is in Black list and White list respectively. If it is present in blacklist then phishing mail warning will be given to the user. If it is present in white list then it is considered as legitimate mail. If it is not present in blacklist or white list then it calls pattern matching module. This module is implemented using AnalyzeDNS algorithm.

As a part of lookup system we are maintaining blacklist and white list. Black list stores list of known fake DNS, while the white list contains list known valid or registered DNS.DNS analyzer uses these list for checking whether the domain in visual link is present in black list or white list. Lookup systems typically have high precision since they are less likely to consider authentic sites as fake. They are also easier to implement than classifier systems. However, lookup systems are more susceptible to higher levels of false negatives .we are maintaining list of blacklisted domains as well as legitimate domain.

Classifier system is used to analyze mail based on heuristic features like URL features from the link, email body features, content features, email header features etc. In proposed classifier system mainly we are using URL features and textual features for performing heuristic analysis of mail.

AS given in reference [15] URL analysis plays very important role in phishing mail detection. We are using 7 features from URL feature and LUA value for lexical URL analysis. As per study in [16] empirical evaluation for feature selection following feature has maximum weight. So we have selected these 6 features for our proposed method.

B. Phishing Mail Detection Algorithm

Phishing mail detection works by analyzing the blacklist and white list checking, senders domain and domain from visual link, textual analysis and lexical analysis. There are different techniques to detect phishing emails that uses email hyperlink properties, e-mail header analysis, file attachment scanning etc. We have developed a algorithm to detect phishing emails. In this algorithm we have used textual analysis and lexical URL analysis.

```
r2=analyzetext(emailtext)
For every link in the email following  Algorithm Executes:
PhishmailDetect ()
{
v_dsn=GetDNS(v_link)
from_dns=GetDNS(From)
if(from_dns exists in blacklist)
    return phishing
else if (from_dns exists in whitelist)
return not phishing
else
r1=analyzeDNS(vlink , fromdns)
r3=analyzelexicalurl(vlink)
if((r1==0)&&(r3==0)&&(r2==0))
print  'Not Phishing'
Else
if(((r2==1)&&(r3==1))||((r1==1)&&(r3==1))||((r2==1)
       &&(r1==1)))
print 'phishing'
 else
print  'possible phishing'
}
```

AnalyzeDNS Algorithm:

Analyze the actual DNS whether it is black listed or white listed or if unknown. Depending on the result it gives warning message.If it is not present in blacklist or white list then it will check for pattern matching. Pattern matching algorithm compares sender dns and actual dns..

Analyze_text Algorithm

Analyzes text and keeps track of number of blacklisted tokens from the mail contents if number of blacklisted tokens present are more than threshold then it is considered phishing and returns true .

Similarly we analyze URL with respect to different features

DOI: 01.IJRTET.8.1.61

and try to analyze links embedded in the email are phishing links or legitimate links.

```
boolean Analyzetext(emailtext)
{
 Set tcount=0
 For(every token in email)
 If (token belongs to blacklisted tokens)
  tcount=tcount+1
 If(tcount >threshold)
          Return  true
  Else
          Return false
}
```

From literature survey we have understood that following are the features which plays very important role in identifying phishing URL. Analyze URL algorithm converts link URL into tokens and keeps count of presence of black listed features from URL. If count is greater than threshold then it returns true else it return false.

```
boolean AnalyzeURL(url)
{
   Convert url into set of tokens
  if (noof@symbol>0)
   count=count+1
  if (noofwordsllinks>0)
   count=count+1
  if (noofperiods>2)
   count=count+1
  if (noofdomains>2)
   count=count+1
  if(length(hostname)>22)
   count=count+1
  if(number_ of_ dash(hostname)>2)
  if (URL contains IP address)
   count=count+1
 if (count> threshold)
          Return true
  else
          Return false
  }
```

IV. Performance Analysis

For phishing e-mail detection we are using data set that is collected from phishing corpus[17].For detection of phishing email we have used domain analysis, textual analysis and URL analysis. We assigned the legitimate mails with 'Positive' answers (P) and the phishing mails with 'Negative' answers (N). True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) can be summarized as below:

TP - legitimate email correctly classified as legitimate email.
FP - legitimate email incorrectly classified as phishing email.
TN - phishing email correctly classified as phishing email.
FN - phishing email incorrectly classified as legitimate email.

Accuracy is the rate of the email correctly classified .False Positives Rate (fpr) measures the rate of legitimate instances that are incorrectly detected as phishing attacks in relation to all existing legitimate instances.

Recall (r) measures the rate of correctly detected phishing attacks in relation to all existing phishing attacks. Precision (p) measures the rate of correctly detected phishing attacks in relation to all instances that were detected as phishing, rate of true negative .Performance of proposed method is analyzed with respect to parameters mentioned above. We have used data set of phishing mails and legitimate mails to check false positive, accuracy, precision and recall.

To evaluate the efficiency of the proposed method, we have applied the proposed algorithm on the data set of phishing emails and legitimate emails[17].We have tested algorithm on 3000 phishing emails from data set. Proposed method Phishmaildetect gives false positive rate of 0.01% and accuracy, precision and recall 99%.

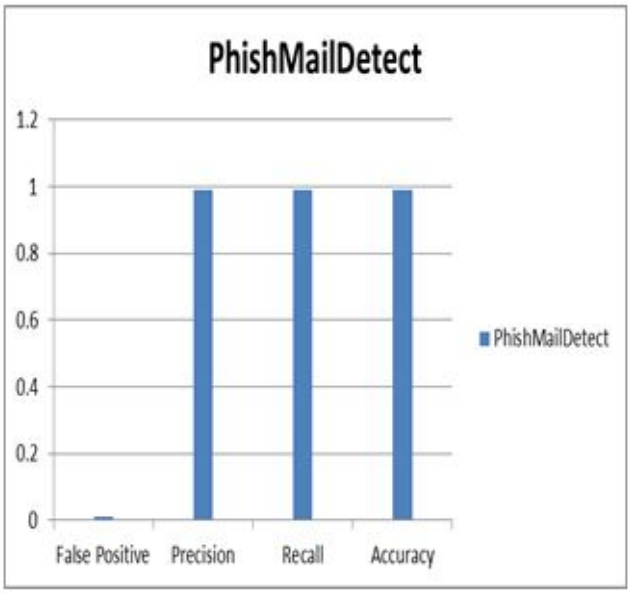

Figure 5 Analysis of proposed method

Figure 5 shows performance analysis of proposed method. Proposed method reduces false positive rate and increases accuracy of detection.

Phishmaildetect technique uses combination of DNS analysis, text analysis and lexical URL analysis. We have compared accuracy of detecting phishing mails using only text analysis, lexical URL analysis with phishmaildetect. We found that accuracy of phishmaildect is better than textual analysis and lexical URL analysis. Comparison of accuracy of phishing mail detection techniques with textual analysis and lexical URL analysis is depicted in figure 6.

V. Performance Comparison

The result of proposed algorithm is compared with the phish-catch and phish-block phishing mail detection methods. From figure 7 and figure 8 it is clear that the proposed method reduces false positive rate and increases accuracy of detection. Phish-catch algorithm has a false positive rate 0.01% and accuracy 84%. Phish block algorithm has a false positive rate 0.1% and accuracy 95%. Proposed method Phishmaildetect gives false positive rate of 0.01% and accuracy 99%. As compared to phish catch and phish block

DOI: 01.IJRTET.8.1.61

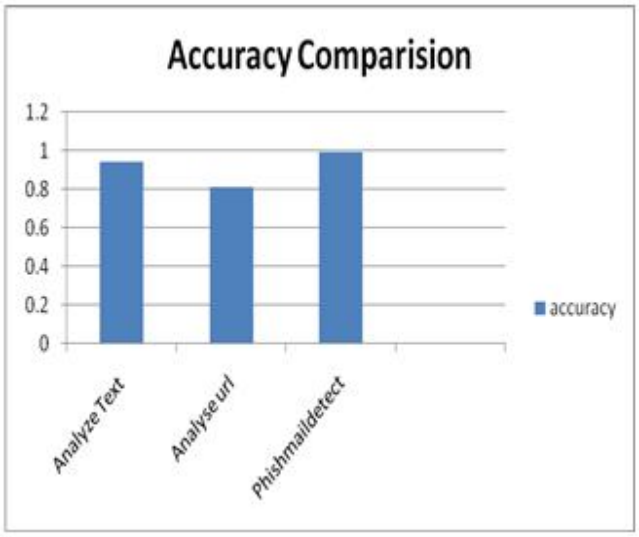

Figure 6. Comparison of textual analysis, URL analysis and phishing mail detection techniques

proposed method improved accuracy as well as false positive rate is also reduced.

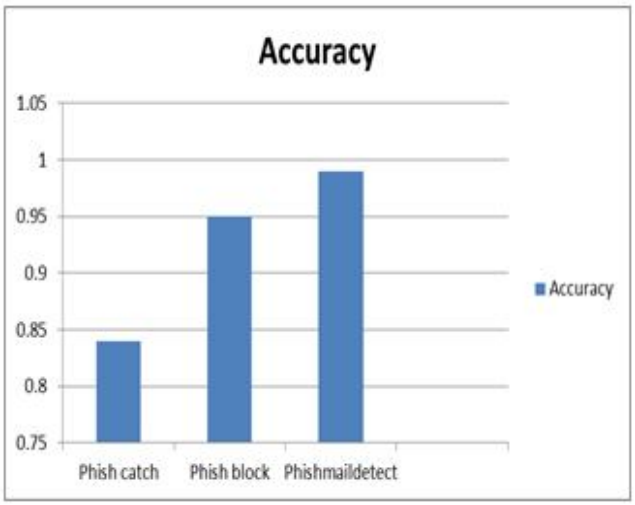

Figure 7.Accuracy comparison of proposed algorithm with Phishblock and phishcatch

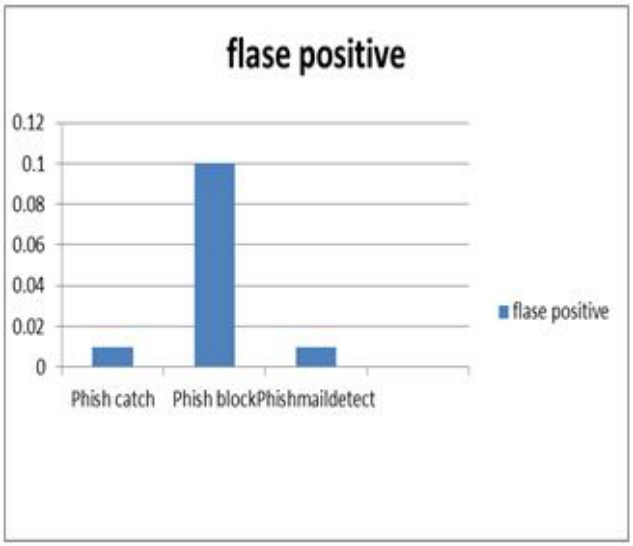

Figure 8 False positive rate comparison

VI. Conclusion

A hybrid method has been proposed to detect phishing mail which is a combination of blacklist, white list and heuristic method. In heuristic detection technique we have considered textual analysis of email and lexical analysis of email for detection. This mechanism effectively detects phishing mails as compared to the previous methods. This mechanism uses combination of textual analysis and lexical URL analysis. From previous study and after analyzing phishing mails it is understood that most of the phishing mails has similar text. So with the help of textual analysis we can effectively determine phishing mail. For increasing effectiveness of mechanism we have used lexical URL analysis. Our main aim was to reduce false positive rate. So by analyzing DNS from the link, textual contents of mail and URL analysis we are trying to reduce false positive rate. At the same time we are taking care of possibility of phishing email then it alerts user with possible phishing. A hybrid method has been proposed and implemented to detect phishing mail which is a combination of blacklist, white list and heuristic method. In heuristic detection technique we have considered textual analysis of email and lexical URL analysis of email for detection. We have used the following steps for implementing our proposed method:

1.Lookup tables are maintained for storing blacklisted domains and legitimate domains.
2. Feature set are identified for detecting phishing mail.
3. Collected phishing mail dataset.
4. Phishing mails are tested with the help of phishmailDetect approach.

We have observed that this mechanism can effectively detect phishing mails as compared to the previous methods. We have evaluated our proposed algorithm and compared it with phish-catch and phish block methods with respect to following performance parameter:

- Accuracy
- False Positive rate
- Precision
- Recall

Proposed method improved accuracy and false positive rate as compared to other methods like phish-catch and phish-block.

References

[1] Danesh Irani, Steve Webb, Jonathon Giffin and Calton Pu, “Evolutionary Study of Phishing”, IEEE International Conference on Web Security, pp. 206-210, 2008.

[2] Cynthia Dhinakaran and Jae Kwang lee, “Reminder: please update your details”, Phishing Trends IEEE first International Conference on Networks & Communications, pp. 295-300, 2009.

[3] Jasveer Singh, “Detection of phishing emails”, International Journal of Computer Science and Technology - IJCST, Vol.2, Issue 3, pp. 547-549, September 2011.

[4] Huajun Huang, Shaohong Zhong, Junshan Tan, “Browser-side Countermeasures for Deceptive Phishing Attack” ,IEEE fifth

DOI: 01.IJRTET.8.1.61

International Conference on Information Assurance and Security, pp.352-355,2009.

[5] A. Alnajim and M. Munro, "An evaluation of user's anti-phishing knowledge retention", IEEE International conference on Information Management and ICIME '09, pp. 210-214,April 2009.

[6] S. Sheng , B. Wardman ,G. Warner , L.F. Cranor , J. Hong and C. Zhang.. "An empirical analysis of phishing blacklists", Sixth International Conference on Email and AntiSpam, July 16-17,2009.

[7] Hossom ,M. A. Fahmy and salma A. Ghoneim , "PhishBlock: A hybrid anti-phishing tool", International Conference on Communications, Computing and Control Applications, IEEE Digital Library, pp. 1-5, March 2011.

[8] Nargundkar S. and Trithani N. ,"Phishcatch: Phishing Detection tool" , 33rd IEEE International Conference on Computer Software and Application, pp.451-456, 2009.

[9] John Yearwood, Musa Mammadov and Arunaya Bannerjee ,""Profiling Phishing Emails Based on Hyperlink Information", International Conference on Advances in Social Networks Analysis and Mining, pp. 1-10, 2010.

[10] Fettee N. Sadeh and A. Tomasic ," Learning to detect Phishing email", Proceedings of the 16th international conference on World Wide Web, Published in ACM digital library, pp.649–656, New York, USA 2007.

[11] A. Bergholz , J. De Beer , S. Glahn , M.F. Moens, Gerhard P. P.,and S. Strobel, "New filtering approaches for phishing email", Journal of Computer Security, vol. 18, pp. 7–35,January 2010.

[12] Jeong-Ho Chang, "Improved Phishing Detection using Model-Based Features", IEEE First International Conference on Networks & Communications, pp 295-300, 2009.

[13] Chandrasekaran, M., Narayanan, K., and Upadhyaya, S., "Phishing E-mail Detection Based on Structural Properties", New York State Cybersecurity Conference Symposium on Information Assurance: Intrusion Detection and Prevention, pp. 2-8. 2006.

[14] Mahmoud Khonji and Youssef Iraqi, "Lexical URL analysis for discriminating phishing and legitimate email", 6th IEEE International Conference on Internet Technology and Secure Transaction, pp.422-427, 2011.

[15] Ma, J., Saul, L., Savage, S., and Voelker, G., "Identifying Suspicious URLs: An Application of Large-Scale Online Learning", Proceedings of the 26th International Conference on Machine Learning, Montreal, Canada, 2009

[16] Fergus Toolan and Joe Carthy, "Feature Selection for Spam and Phishing Detection", IEEE International Conference eCrime Researchers Summit, pp.1-12, 2010.

[17] J .Nazrio ,"Phishing corpus, "http://monkey.org /_jose/wiki/doku.php? id=phishingcorpus, accessed July 2010.

DOI: 01.IJRTET.8.1.61